# Web Scraping em R

Anthony Daniel – 222027786
Bruna Analhys – 150057822
Enzo Seibel Fasano – 222015275
João Vitor Rodrigues – 200064835

# Roteiro

1) Definição e aplicações
2) Quando usar/não-usar web scraping
3) Cuidados necessários
4) Web Scraping no R
5) Exemplo 1: Impostômetro
6) Introdução à HTML
7) SelectorGadget
8) Exemplo 2: IMDB
9) Exemplo 3: TripAdvisor
10) Comentários Finais

# O que é Web Scraping?

Web Scraping é um metodo de "raspagem" de dados de sites que usa scripts para obter as informações necessárias, simulando um comportamento "humano".

# O que é Web Scraping?

Websites with HTML Pages

Web Scraping Technology


Structured Data

Figura. Visao geral de um *web scraping.*

# Aplicações do Web Scraping


Figura. Aplicações do *web scraping*.

# Quando usar Web Scraping

- Quando os dados que você precisa estão disponíveis em um site, mas não em um formato fácil de baixar
- Quando precisamos coletar um volume grande de dados da internet

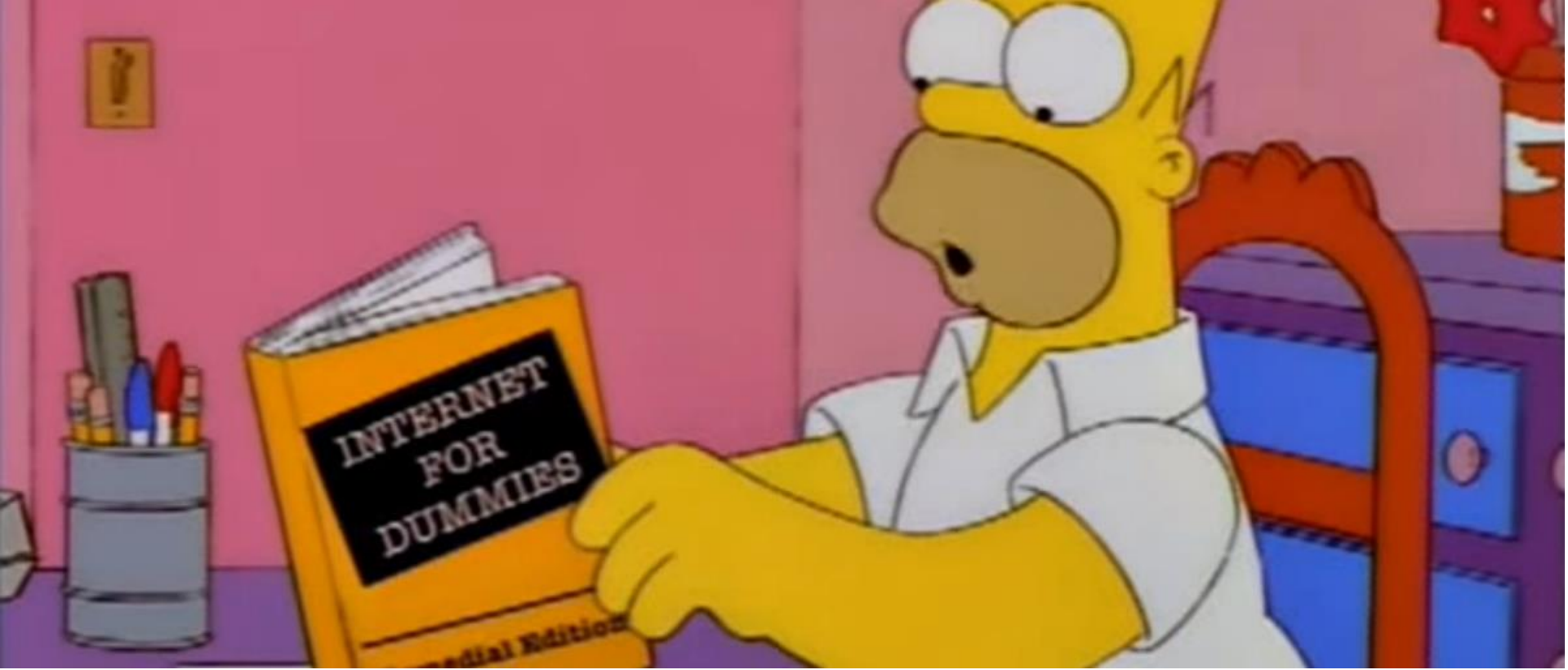
# Quando **não** usar Web Scraping

- Quando temos uma forma mais simples de obter os dados (API, base de dados, etc.)
- Quando os termos de uso do site não nos permitem fazer isso
- Quando houver risco de derrubar ou comprometer a estabilidade do site
- Quando as informações do site não são públicas

# Precauções ao usar Web Scraping

- É crucial respeitar a Lei Geral de Proteção de Dados –LGPD

- Verificar a propriedade dos dados e usar ferramentas como o pacote *polite* e a extensão *robots.txt* para garantir que você está coletando dados de maneira ética e responsável.




"Be polite"! Um scraping pode sobrecarregar um servidor, principalmente, se o script estiver fazendo uma grande quantidade de solicitações. Respeite o robots.txt!

# Noções sobre página da Web

- Quando uma pagina na Internet é visitada, o navegador faz uma solicitação a um servidor web. Essa solicitação é chamada de GET, pois são recebidos arquivos do servidor.
  - HTML: contem o conteudo principal da pagina.
  - CSS: adiciona estilos para que a pagina fique customizada.
  - JS: arquivos JavaScript adicionam interatividade a pagina.

# Noções sobre página da Web

**HTML**
structure

**CSS**
presentiitJon/1ppearanc:e

**JavaScript**
dyn1mlsm/1ctlon

Figura. Camadas do desenvolvimento *web*.

# Web Scraping em R

- Existem vários pacotes em R que são frequentemente utilizados para fazer Web Scraping. Alguns dos mais usados são:
  - XML2
  - Rselenium
  - **Rvest**

# Pacote rvest

- É uma das principais bibliotecas em R para Web Scraping. Ele foi desenvolvido por Hadley Wickham e oferece uma série de funções fáceis de usar para interagir com páginas HTML.

# Pacote rvest

Principais funções do rvest:

- read_html(): Essa função é usada para ler o conteúdo HTML de uma página da web e armazená-lo em um objeto que pode ser manipulado.
- html_table(): Essa função permite extrair tabelas de uma página da web. Ela é útil quando você deseja obter os dados estruturados em uma tabela diretamente em um data frame no R.
- html_nodes(): Com essa função, você pode selecionar elementos HTML específicos com base em seus seletores CSS. Ela retorna uma lista contendo todos os elementos correspondentes.
- html_text(): Essa função permite extrair o texto de um ou mais elementos HTML selecionados.
- html_attr(): Usada para extrair atributos específicos de elementos HTML. Com ela, você pode obter valores de atributos como IDs, classes, URLs de links, caminhos de imagens, entre outros.

# Exemplo 1: Impostômetro

# Selector Gadget

- É uma extensão do Chrome de código aberto e fácil de usar que pode ajudar você a encontrar o seletor certo para qualquer página da Web.

# Selector Gadget

## Data Structures in R Programming

A data structure is a particular way of organizing data in a computer so that it can be used effectively. The idea is to reduce the space and time complexities of different tasks. Data structures in R programming are tools for holding multiple values.

R's base data structures are often organized by their dimensionality (1D, 2D, or nD) and whether they're homogeneous (all elements must be of the identical type) or heterogeneous (the elements are often of various types). This gives rise to the five data types which are most frequently utilized in data analysis. the subsequent table shows a transparent cut view of those data structures.

| DIMENSION | HOMOGENOUS | HETEROGENEOUS |
|---|---|---|
| 1D | Vector | List |
| 2D | Matrix | Dataframe |
| nD | Array | |

p | Clear (111) | Toggle Position | XPath | ? | X


# Exemplo 2: IMDB


# Exemplo 3: TripAdvisor


O Web Scraping desempenha um papel fundamental nas análises estatísticas e de mercado, fornecendo uma fonte rica e diversificada de dados para a tomada de decisões e pesquisas em várias áreas

# OBRIGADO!

Alguma dúvida?